# SALES PREDICTION

(Unleashing Predictive Insights with Big Mart Sales Forecasting)

## A MINI-PROJECT REPORT

***Submitted by***

**PRIYADHARSHINI G (210701198)**

**SHREENIDHI G L (210701246)**

***in partial fulfillment of the award of the degree***

***of***

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

RAJALAKSHMI
ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

**RAJALAKSHMI ENGINEERING COLLEGE, THANDALAM**

**MAY - 2024**

## BONAFIDE CERTIFICATE

Certified that this project **“SALES PREDICTION”** is the bonafide work of **“PRIYADHARSHINI G 210701198”** and **“SHREENIDHI G L 210701246”** who carried out the project work under my supervision.

**SIGNATURE**

**Dr.RAKESH KUMAR M ,**

Assistant Professor,

Computer Science & Engineering

Rajalakshmi Engineering College

(Autonomous)

Thandalam,Chennai-602105

This mini project report is submitted for the viva voce examination to be held on __________

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

# ACKNOWLEDGEMENT

We express our sincere thanks to our beloved and honorable chairman **MR. S. MEGANATHAN** and the chairperson **DR. M. THANGAM MEGANATHAN** for their timely support and encouragement.

We are greatly indebted to our respected and honorable principal **Dr. S.N. MURUGESAN** for his able support and guidance

No words of gratitude will suffice for the unquestioning support extended to us by our Head of The Department **Dr.P. KUMAR M.E Ph.D.,** and our Academic Head **Dr.N.DURAIMURUGAN,** for being ever supporting force during our project work.

We also extend our sincere and hearty thanks to our internal guide **Dr. M. RAKESH KUMAR, M.E Ph.D.**, for her valuable guidance and motivation during the completion of this project.

Our sincere thanks to our family members, friends and other staff members of computer science engineering.

Priyadharshini G (210701198)

Shreenidhi G L (210701246)

## TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# ABSTRACT

The traditional way of finding the sales and marketing goals no longer help the companies to increase the pace of the sales of the companies among the competitive market. As these approaches do not have the insights to customers' purchasing patterns. The main issue faced by retailers is variations in sales of a product. In order to address this challenge, we aim to forecast sales by analyzing historical sales data across various stores. Thanks to these advancements, crucial aspects like consumer purchasing trends, target demographics, and forecasting sales for the upcoming years can now be easily discerned, aiding the sales team in crafting strategies to enhance business growth. This paper aims to introduce an application for predicting future sales of stores, leveraging insights from previous years' sales data. A thorough examination of sales prediction is conducted using Machine Learning models using XGBoost Regressor. Key predictive parameters encompass item weight, item fat content, item visibility, item type, item MRP, outlet establishment year, outlet size, and outlet location type.

# CHAPTER 1

# INTRODUCTION

## 1.1 INTRODUCTION

The Sales prediction is crucial for businesses, enabling them to increase profits by implementing effective sales strategies. Accurate forecasts rely on analyzing various factors such as customer preferences, purchasing patterns, and the shop's environment, using historical sales data. This project addresses store sales prediction by dividing the process into phases, starting with data preprocessing. The dataset is imported, examined for missing values, and processed using statistical methods. Feature engineering uncovers additional insights, preparing the data for model training. The dataset is then split into training and testing sets using Scikit-learn, ensuring a proper ratio to prevent overfitting and underfitting. By analyzing past sales data with these methods, businesses can optimize operations and enhance decision-making, ultimately boosting sales performance.

## 1.2 SCOPE OF THE WORK

The scope of work for a sales prediction project encompasses collecting and cleaning historical sales data, performing feature engineering to extract meaningful insights, conducting exploratory data analysis to understand sales trends and patterns, selecting and training machine learning models for sales forecasting, evaluating model performance using appropriate metrics, fine-tuning models for optimal results, generating sales forecasts for future periods, interpreting results, and presenting findings with actionable recommendations for business decisions. Implementation of the models into the business workflow and continuous monitoring and updating of the models for accuracy and relevance are also key aspects of the project scope

## 1.3 PROBLEM STATEMENT

The problem statement revolves around the challenge of inaccurate sales predictions due to data complexity, quality issues, inefficient sales strategies, market volatility, and limited scalability and adaptability of existing models. Businesses struggle with suboptimal inventory management, pricing strategies, and decision-making processes, leading to financial losses, customer dissatisfaction, and competitive disadvantages.There is a pressing need to develop robust, data-driven sales prediction models that can handle diverse data sources, address data quality issues, consider external market factors, and provide scalable solutions for improved sales forecasting and strategic decision-making in dynamic business environments.

## 1.4 AIM AND OBJECTIVES OF THE PROJECT

The aim of this project is to develop a robust sales prediction model that leverages data-driven insights to accurately forecast future sales for a business, addressing challenges such as data complexity, quality issues, inefficient sales strategies, market volatility, and limited model scalability. The objectives include collecting and preprocessing historical sales data, performing feature engineering and selection, conducting exploratory data analysis for insights, building and evaluating machine learning models, optimizing and validating model performance, integrating the model into business workflows, and ensuring continuous monitoring, adaptation, and improvement to support informed decision-making and enhance sales forecasting accuracy and effectiveness.

# CHAPTER 2

# SYSTEM SPECIFICATIONS

## 2.1 HARDWARE SPECIFICATIONS

| | | |
|---|---|---|
| Processor | : | Intel i3 |
| Memory Size | : | 8GB (Minimum) |
| HDD | : | 1 TB (Minimum) |

## 2.2 SOFTWARE SPECIFICATIONS

| | | |
|---|---|---|
| Operating System | : | WINDOWS 10 |
| Front – End | : | Html,CSS,Javascript,Fast API |
| Back - End | : | Python |

# CHAPTER 3

## MODULE DESCRIPTION

The Big Mart Sales Prediction module is designed to leverage advanced data analytics and machine learning techniques to accurately forecast sales for Big Mart stores. By analyzing historical sales data and various external factors, this module aims to provide actionable insights and predictions to optimize inventory management, marketing strategies, and overall sales performance.

**Features:**

1. **Data Preprocessing and Cleaning:**
   - Handles missing values and inconsistent data.
   - Normalizes and standardizes data for optimal model performance.
   - Feature engineering to create meaningful variables from raw data.
2. **Exploratory Data Analysis (EDA):**
   - Visualizes sales trends and patterns over time.
   - Identifies key factors influencing sales.
   - Generates insights on product performance and customer behavior.
3. **Sales Prediction Models:**
   - Implements multiple machine learning algorithms using XGBoost.
   - Utilizes time series analysis for more accurate short-term and long-term forecasting.
4. **Model Evaluation and Validation:**
   - Uses metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared to evaluate model performance.
   - Provides visualizations for actual vs. predicted sales to assess model accuracy.
5. **Deployment and Scalability:**
   - Offers deployment options using Fast APIs for integration with existing systems.
   - Scalable to handle large datasets and multiple store locations.

**Benefits:**

- **Improved Inventory Management:** Predicts demand more accurately to reduce overstock and stockouts.
- **Enhanced Marketing Strategies:** Identifies trends and seasonal effects to optimize promotional activities.
- **Data-Driven Decision Making:** Empowers stakeholders with reliable forecasts to make informed decisions.

- **Increased Sales Performance:** Helps in aligning supply chain and operational strategies to meet customer demand efficiently.

**Use Cases:**

- Retail managers seeking to optimize store operations.
- Marketing teams planning targeted campaigns.
- Supply chain professionals looking to improve logistics and reduce costs.
- Data analysts and scientists aiming to develop robust predictive models for retail data.

The Big Mart Sales Prediction module is an essential tool for any retail business looking to leverage data for enhanced decision-making and competitive advantage.

# CHAPTER 4

# SYSTEM DESIGN

## 4.1 SYSTEM ARCHITECTURE DIAGRAM

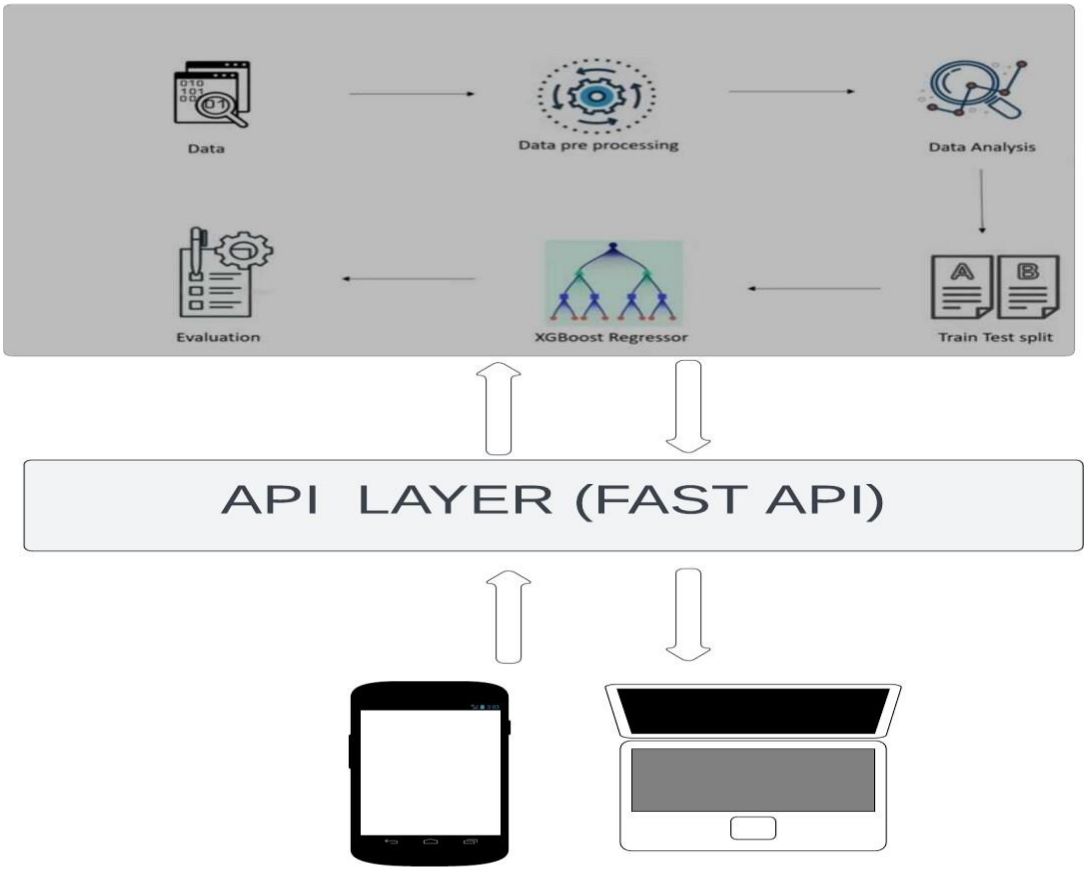


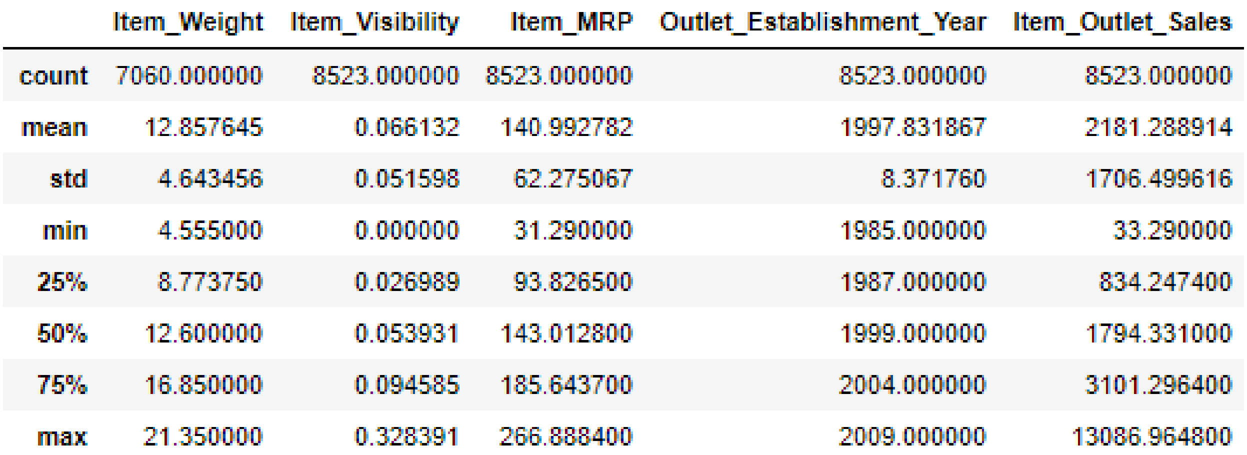
Fig 1 Shows the architecture diagram of Sales prediction

The architecture of the Big Mart Sales Prediction module integrates XGBoost and FastAPI to deliver a robust and efficient forecasting solution. The process starts with Data Ingestion, collecting historical sales data and relevant external factors. This data is then meticulously preprocessed and cleaned, handling missing values and standardizing features. Exploratory Data Analysis (EDA) uncovers patterns and trends, informing the Feature Engineering phase to enhance predictive power. The core Model Training phase employs XGBoost for its exceptional performance and scalability, fine-tuned through hyperparameter tuning and cross-validation. Model Evaluation ensures accuracy using metrics like RMSE. The trained model is seamlessly deployed using FastAPI, enabling real-time predictions and integration with front-end applications.

# CHAPTER 5

# TABLES

Table 5.1: Sales Data



| | Item_Weight | Item_Visibility | Item_MRP | Outlet_Establishment_Year | Item_Outlet_Sales |
|---|---|---|---|---|---|
| **count** | 7060.000000 | 8523.000000 | 8523.000000 | 8523.000000 | 8523.000000 |
| **mean** | 12.857645 | 0.066132 | 140.992782 | 1997.831867 | 2181.288914 |
| **std** | 4.643456 | 0.051598 | 62.275067 | 8.371760 | 1706.499616 |
| **min** | 4.555000 | 0.000000 | 31.290000 | 1985.000000 | 33.290000 |
| **25%** | 8.773750 | 0.026989 | 93.826500 | 1987.000000 | 834.247400 |
| **50%** | 12.600000 | 0.053931 | 143.012800 | 1999.000000 | 1794.331000 |
| **75%** | 16.850000 | 0.094585 | 185.643700 | 2004.000000 | 3101.296400 |
| **max** | 21.350000 | 0.328391 | 266.888400 | 2009.000000 | 13086.964800 |

## CHAPTER 6
## SAMPLE CODING

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from xgboost import XGBRegressor
from sklearn import metrics
# loading the data from csv file to Pandas DataFrame
big_mart_data = pd.read_csv('/content/Train.csv')
# first 5 rows of the dataframe
big_mart_data.head()
# number of data points & number of features
big_mart_data.shape
# getting some information about thye dataset
big_mart_data.info()
# checking for missing values
big_mart_data.isnull().sum()
# mean value of "Item_Weight" column
big_mart_data['Item_Weight'].mean()
# filling the missing values in "Item_weight column" with "Mean" value
big_mart_data['Item_Weight'].fillna(big_mart_data['Item_Weight'].mean(), inplace=True)
# mode of "Outlet_Size" column
big_mart_data['Outlet_Size'].mode()
# filling the missing values in "Outlet_Size" column with Mode
mode_of_Outlet_size = big_mart_data.pivot_table(values='Outlet_Size', columns='Outlet_Type',
aggfunc=(lambda x: x.mode()[0]))
print(mode_of_Outlet_size)
miss_values = big_mart_data['Outlet_Size'].isnull()
big_mart_data.loc[miss_values, 'Outlet_Size'] =
big_mart_data.loc[miss_values,'Outlet_Type'].apply(lambda x: mode_of_Outlet_size[x])
# checking for missing values
```

```
big_mart_data.isnull().sum()
# Item_Type column
plt.figure(figsize=(30,6))
sns.countplot(x='Item_Type', data=big_mart_data)
plt.show()
big_mart_data['Item_Identifier'] = encoder.fit_transform(big_mart_data['Item_Identifier'])
big_mart_data['Item_Fat_Content'] = encoder.fit_transform(big_mart_data['Item_Fat_Content'])
big_mart_data['Item_Type'] = encoder.fit_transform(big_mart_data['Item_Type'])
big_mart_data['Outlet_Identifier'] = encoder.fit_transform(big_mart_data['Outlet_Identifier'])
big_mart_data['Outlet_Size'] = encoder.fit_transform(big_mart_data['Outlet_Size'])
big_mart_data['Outlet_Location_Type'] =
encoder.fit_transform(big_mart_data['Outlet_Location_Type'])
big_mart_data['Outlet_Type'] = encoder.fit_transform(big_mart_data['Outlet_Type'])
X = big_mart_data.drop(columns='Item_Outlet_Sales', axis=1)
Y = big_mart_data['Item_Outlet_Sales'
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
# prediction on training data
training_data_prediction = regressor.predict(X_train)
# R squared Value
r2_train = metrics.r2_score(Y_train, training_data_prediction)
print('R Squared value = ', r2_train)
# prediction on test data
test_data_prediction = regressor.predict(X_test)
# R squared Value
r2_test = metrics.r2_score(Y_test, test_data_prediction)
print('R Squared value = ', r2_test)
```

# CHAPTER 7

# SCREENSHOTS

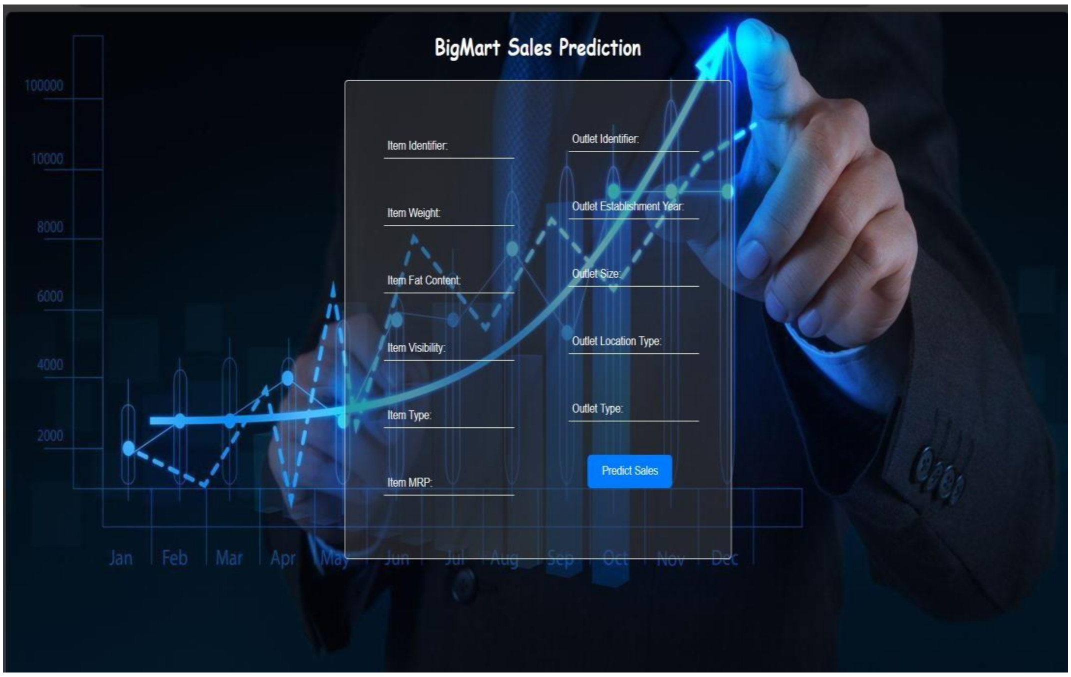
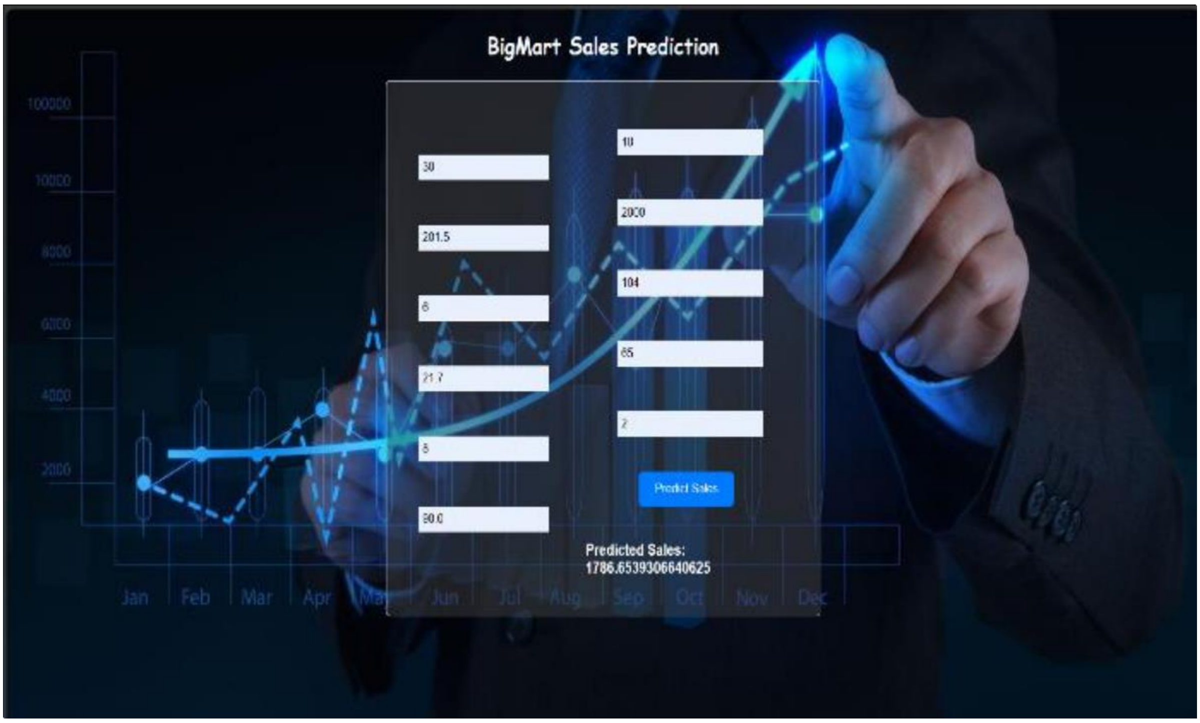


Fig 2 Shows the home page of the sales prediction

The Fig 2 shows the output screenshot of the home page of sales prediction app where the user needs to enter the details of item identifier item weight item fat content ,item visibility ,item type ,outlet establishment year, outlet size , outlet location type ,outlet type.

Fig 3 shows the predicted sales of the item 30 after entering the details of item identifier item weight item fat content ,item visibility ,item type ,outlet establishment year, outlet size , outlet location type ,outlet type.



BigMart Sales Prediction

30

201.5

6

21.7

8

80.0

10

2000

104

65

2

Predicted Sales:
1786.6539306640625

Fig 3 Shows the predicted sales

# CHAPTER 8
# CONCLUSION AND FUTURE ENHANCEMENT

In concluding this project, it is evident that Machine Learning (ML) has become an indispensable tool for shaping contemporary business strategies, especially in understanding and leveraging consumer purchase patterns. Traditional methodologies, while valuable, often fall short in driving substantial revenue growth. ML's predictive capabilities, fueled by vast datasets and sophisticated algorithms, empower businesses to anticipate market dynamics with precision. By analyzing historical sales data alongside factors such as market trends, customer demographics, and promotional activities, ML-driven sales prediction not only forecasts future sales trajectories but also guides strategic decision-making. This transformative approach enables organizations to optimize inventory management, allocate resources effectively, and tailor marketing strategies to target specific customer segments, thus enhancing overall competitiveness and market responsiveness.

As businesses navigate an increasingly dynamic and competitive landscape, the adoption of ML for sales prediction emerges as a strategic imperative. Beyond numerical forecasts, ML offers actionable insights that pave the way for innovation, customer-centricity, and sustainable growth. By embracing ML's predictive prowess, organizations position themselves to thrive amidst uncertainties, capitalize on emerging opportunities, and cultivate enduring customer relationships. As we look towards the future, the integration of ML into business strategies not only shapes revenue growth but also fosters agility, resilience, and strategic foresight, ensuring sustained success in an ever-evolving marketplace.

# REFERENCES

[1] L. Bornmann and H. D. Daniel, welches in:We know that h index is a measure of the performance of a journal or a researcher over a given time frame. Jasmin, K. (2007) The impact of personalized email campaigns on e-mail list rental and co-registration. Journal of the American Society for Information Science & Technology, 58 (9), 1381-1385.

[2] Bornmann L, & Daniel H. D. (2009). The state of h index research: In the light of the above discussions, it can be posited that while the h index offers a more accurate depiction on the productivity of its subject of study than the ISI, it may not be the perfect way of measuring the research performance of individuals or institutions. EMBO reports, 10(1), 2-6.

[3] Carpenter, M. P. , & Narin F. Predictors of freshman course difficulty: Cross sectional data. ISSN: 0005-9360, Information Sciences The adequacy of the science citation index as an indicator of international scientific activity. Information processing & management, 20 (4-5), 447-454.

[4] Dašić P, Moldovan L, Grama L Effect of Protein Kinase C on Fatty Acid Composition in Adipose Tissue of Diabetic Rats. Publication analysis of research papers emerging from Romanian and Serbian institutions in the SCI, SCI-E and SSCI citation indices. Procedia Technology, 19, 1075-1082.

[5] Egghe, L. (2006). G-index and complementary concepts in the organizational framework. Scientometrics, 69(1), 131-152. [6] Egghe, L. (2006). An improvement of the h-index: The g-index The h-index is a measure of an individual's academic and research performance, which is integral for academic promotion and tenure assessments worldwide, particularly in the United States of America. ISSI.

[6] Garfield, E. (2007). It returned back a list of the evolution of science citation indexes in the university of science technology. International microbiology, 10(1), 65.

[7] Hirsch, J. E. (2005). To tier or not to tier?: A synthesis and evaluation of a decade of research on citation practices by humanities scholars. Academicus: The International Scientific Journal , 6, 5 – 27. According to the given context, it can be defined as: “The term that has been used to create a measure of productivity of a scientist”. PNAS. 102 (46): 19849 19852.

[8] Jacsó, P. (2008). CASH HAMMOND.It is evident that the availability of knowledge resources has increased with the rise of internet usage, as mentioned in an article published in the Online Information Review where it highlighted that knowledge resources are accessible by 32(4), 524–535.